An Exploratory Data Analysis on Air Quality Data of Trivandrum

V. Santhana Lakshmi and M. S. Vijaya

Abstract Data analysis is the most integral part of any research. It is the process of examining the data using statistical methods to identify the hidden patterns and trends which aid in making decisions. This helps in understanding the distribution, correlation, outliers, and missing values found in the data. In this paper, data analysis is performed over the air pollutant data and the meteorological data that influences air pollution. The meteorological data for the period of 4 years of Trivandrum city was taken for the purpose of analysis. The dataset includes 26,544 instances and 23 features. Pollutant parameters such as $PM_{2.5}$, PM_{10}, CO, SO_2, ozone, NO_X, and NH_3 are considered for analysis. Meteorological features taken for analysis include temperature, dew, humidity, wind speed, wind direction, etc. Meteorological features play a substantial role in identifying air pollution. Boxplots, heat maps, pair plots, and histograms were used to reveal the distribution and correlation between the attributes. From the analysis, it has been identified that the features like sea level pressure, $PM_{2.5}$, PM_{10}, CO, NO_X, NH_3, SO_2, and ozone are positively correlated with air quality index whereas features like, dew, humidity, wind speed, cloud cover are negatively correlated with air quality index. The results of the data analysis assist in preparing the data for further research.

Keywords Correlation · Outliers · Boxplot · Heat map · Pair plot · Histogram

1 Introduction

Ambient air pollution is becoming a serious hazard to the health of human beings. Clean air gets polluted due to the introduction of chemicals, particulate matter, and biological materials, which are called aerosols that cause harm to humans, other

V. Santhana Lakshmi (✉) · M. S. Vijaya
PSGR Krishnammal College for Women, Peelamedu, Coimbatore, Tamilnadu, India
e-mail: sanlakmphil@gmail.com

M. S. Vijaya
e-mail: msvijaya@psgrkcw.ac.in

A. Joshi et al. (eds.), *Information and Communication Technology for Competitive Strategies (ICTCS 2022)*, Lecture Notes in Networks and Systems 623,
https://doi.org/10.1007/978-981-19-9638-2_68

living organisms, and the natural environment. The size of the aerosol, also known as particulate matter, ranges from 0.001 to 10 m, and it damages the respiratory organs when inhaled by people [1]. Due to the very small size of the aerosol, it effortlessly enters into the respiratory organs. Lung illness was found in people who had been exposed to contaminated air for a long time. Every year, air pollution has a negative impact on the health of 9% of the population.

It is thought to be the most significant risk factor affecting human health. According to a Lancet survey, air pollution was responsible for 167 million deaths in India in 2019, accounting for 178% of the country's total mortality. The majority of these deaths were caused by particulate matter pollution in the environment. The death rate due to ambient particulate matter pollution increased by 115.3% [2].

Particle pollution, ground-level ozone, carbon monoxide, sulphur oxides, nitrogen oxides, and lead are six significant air pollutants that cause serious health and environmental problems, according to the World Health Organization (WHO). Government takes necessary steps to control air pollution. Effective systems that can forecasting the air quality and generate warnings based on the results are therefore required and important for society in providing health alerts when the level exceeds the specific limit. To build a successful forecasting model, extensive understanding of the data is required to identify patterns within the data.

It was also identified that meteorological factors stimulus more on concentration levels of pollutants. Some of the meteorological factors that influence this are wind speed & direction, temperature, humidity, rainfall, and solar radiation. In this paper, exploratory data analysis performed on air quality dataset, which encompasses pollutant features and meteorological features is described, and the statistical characteristics of the data are examined.

2 Significance of EDA

Exploratory data analysis is a process of assessing datasets using statistical graphics and visualization techniques to analyse their essential characteristics, hidden patterns, and trends. Data analysis and data preprocessing are the crucial steps involved in the process of building a machine learning model. Exploratory data analysis is a key and normally the first task to be performed in information mining. It allows us to visualize the data, to apprehend it, as properly as to create hypotheses for analysis. Exploratory data analysis aids in understanding the distribution of the data.

It helps to identify the role of attributes for the purpose of prediction. Real-world data cannot be consumed directly for developing a model since it involves missing values, wrong values, and outliers. Dispersion of the data, correlation between the attributes and outliers can be identified through data analysis. The results of exploratory data analysis help in identifying the attributes that require preprocessing. The findings of EDA will suggest the kinds of data preprocessing such as data cleaning, outlier removal and transformation.

3 Exploratory Data Analysis

Exploratory data analysis is the process of exploring the data through statistical methods, and visualization techniques to identify the hidden patterns, trends in the data. Exploratory data analysis involves four steps which includes problem definition, data preparation, data analysis, and representation of the results [3]. The following section explains the various steps followed for performing exploratory data analysis on air pollution data.

3.1 Problem Definition

Air pollution leads to an expand in the mortality rate. Air gets polluted when strong particles and gases in the form of aerosols enter it. The factors that purpose air pollution can be extensively classified into man-made and natural sources. Natural sources of air pollution consist of dirt from the earth's surface, sea salt, volcano eruptions, and wooded area fires. Man-made sources of air pollution include industrial emissions, transportation emissions, and agriculture. Since air pollution is directly associated with the health of the human being, it is very much essential to build a machine learning model to predict the air quality index. The main objective of the research was to get insights from the raw pollutant data and meteorological data to build a machine learning model to predict the air quality index for the period of three years. Air quality index is a measure that discloses how polluted the air was. The higher the AQI value, the higher the level of air getting polluted. The values lie between 0 and 500. In order to build a predictive model, it is necessary to analyse the role of each and every parameter in performing the prediction.

3.2 Data Collection

Trivandrum has been chosen as the location to predict air quality. It is the largest city in Kerala. The city has a population of 957,730 people, with a population of 1.68 million in the metropolitan area. Thiruvananthapuram district is situated between north latitudes 8º 17′ and 8º 54′ and east longitudes 76º 41′ and 77º 17′. The southernmost extremity, "Parassala," is 56 kms away from Kanyakumari, the "lands end of India." The district stretches along the shores of the Arabian Sea for a distance of 78 kms [4]. The outline of the state is provided in Fig. 1. Air pollutant data has been collected from Central Control Room for Air Quality Management, Central Pollution Control Board, Delhi, through portal [5]. The details such as state, city, station, and duration for which the data is required and parameters required should be provided in the portal. Then, the requested data can be downloaded in the form

of csv. The steps of downloading the data are provided in the form of screenshots in Figs. 2 and 3.

Meteorological data was collected from Visual Crossing Web site [6]. On providing the details such as city and duration, the requested data can be extracted. The screenshot that contains sample data is provided in Fig. 4. Hourly data was taken over a period of three years. The total number of instances used is 26,282. Meteorological parameters used for analysis include temperature, relative humidity, dew, sea level pressure, cloud cover, visibility, conditions, icon, SR, BP, AT, RF, wind speed, and wind direction. Air pollution parameters used for analysis include $PM_{2.5}$, PM_{10}, CO, SO_2, ozone, NO_X, and NH_3.

Fig. 1 Trivandrum

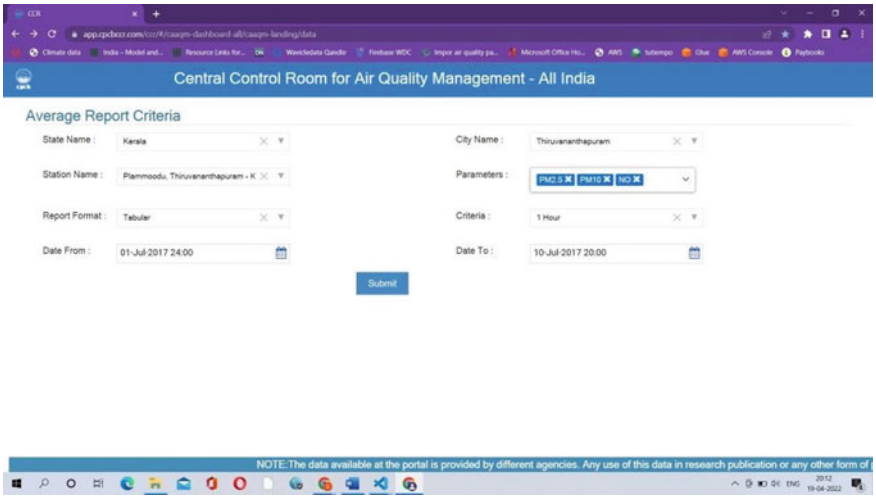

Fig. 2 Central pollution control board portal

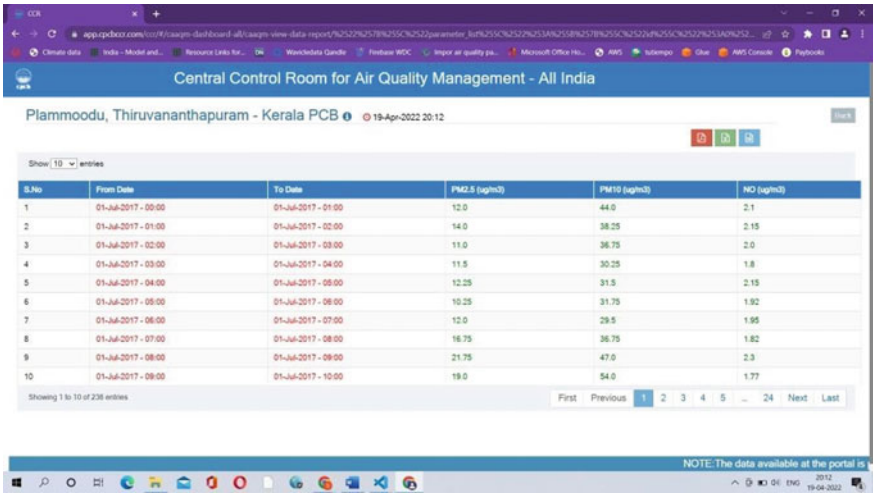

Fig. 3 Sample data from central pollution control board portal

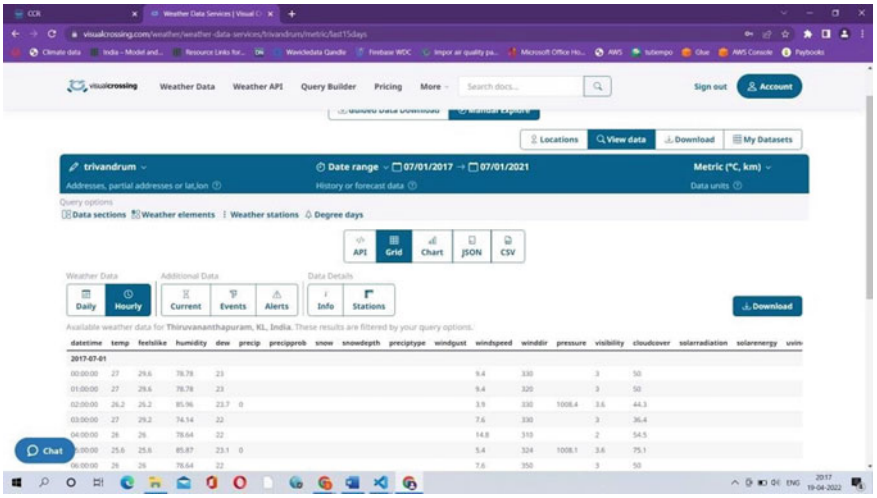

Fig. 4 Sample meteorological data from visual crossing website

The pollutant parameters obtained from pollution control boards for the period of three years are combined with meteorological parameters obtained from the Visual Crossing Web site for the same period using date with timestamp as a common feature.

In depth analysis needs to be performed on the attributes to identify their importance in building the prediction model.

Table 1 Air quality standards

Air quality index(AQI)	Category
0–50	Good
51–100	Satisfactory
101–200	Moderate
201–300	Poor
301–400	Very poor
401–500	Severe

Table 2 Health breakpoints of pollutants

AQI category	PM_{10}	$PM_{2.5}$	NO_2	O_3	CO	SO_2	NH_3
Good (0–50)	0–50	0–30	0–40	0–50	0–1.0	0–40	0–200
Satisfactory (51–100)	51–100	31–60	41–80	51–100	1.1–2.0	41–80	201–400
Moderate (101–200)	101–250	61–90	81–180	101–168	2.1–10	81–380	401–800
Poor (201–300)	251–350	91–120	181–280	169–208	10.1–17	381–800	801–1200
Very poor (301–400)	351–430	121–250	281–400	209–748	17.1–34	801–1600	1201–1800
Severe (401–500)	430 +	250 +	400 +	748 +	34 +	1600 +	1800 +

There are six categories of AQI which include Good, Satisfactory, Moderately Polluted, Poor, Very Poor, and Severe. The AQI category, corresponding ambient concentration and health breakpoints are provided in Tables 1 and 2.

3.3 Dataset Creation

The dataset includes meteorological features extracted from visual crossing Web site and pollutant features extracted from the Central Pollution Control Board portal. Meteorological parameters used for analysis include temperature, relative humidity, dew, sea level pressure, cloud cover, visibility, conditions, icon, SR, BP, AT, RF, wind speed, and wind direction. Air pollution parameters used for analysis include $PM_{2.5}$, PM_{10}, CO, SO_2, ozone, NO_X, and NH_3.

Meteorological Parameters

The meteorological parameters taken for analysis include temperature, due, humidity, sea level pressure, cloud cover, visibility, conditions, icon, relative humidity, barometric pressure, air temperature, solar radiation, wind speed, and direction. Temperature is a measure used to identify how hot or cold the day is. It is a measure of the kinetic energy present in the molecules. The energy is directly proportional to the temperature [7]. The higher the energy, the higher the temperature. It is a measure of how fast air molecules are moving from one place to another. Temperature along with wind speed decides whether pollutants are dispersed or reside in the same location. In addition, temperature, when combined with solar radiation, creates photochemical smog as a result of undergoing chemical reaction.

The feels-like temperature refers to the amount of hotness or coldness felt outside. It is calculated based on parameters such as temperature, humidity, and the speed of the wind. Relative humidity denotes the amount of water vapour present in the air. It is the ratio of the moisture content maintained in the air with the given temperature. Dew is the condensed form of water vapour which is called a droplet. Dew point refers to the temperature at which water vapour condenses to form a droplet. The presence of dew in a surface layer reduces the pollutant concentration. When water vapour condenses to form droplets, it removes the gaseous pollutants in the area, so dew is indirectly proportional to air pollution.

Barometric pressure refers to atmospheric pressure. Sea level pressure represents the thermal difference between sea and land. Atmospheric pressure encountered at any elevation, calculated by a formula, is decreased to a value that approximates the pressure at sea level. Cloud cover is a measure of cloudiness prevailing in the sky. It is measured in terms of Okta. The lightness or darkness of air pollution particles has an effect on cloud formation, according to NASA scientists. This has an effect on the Earth's climate as well.

Visibility is the measure of transparency. It is a measure of the distance at which an object on the ground can be seen clearly. When aerosols are present in the air, they create a white haze which affects the clarity of identifying objects present faraway in the ground. Solar radiation represents the amount of energy emitted by the sun. Rainfall is the total amount of water that falls in an area. Wind speed refers to the speed at which the wind flows from a direction. It has been identified that rainfall and wind speed are indirectly proportional to the air quality index. Rainfall dissolves all the pollutants found in the atmosphere. Wind disperses the pollutant particles from one place to another and causes the pollutant particles to disappear.

Pollutant Parameters

Sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3), carbon monoxide (CO), $PM_{2.5}$, and PM_{10} particulate matter are the six primary air pollutants identified by India's Central Pollution Control Board (CPCB) [8]. The concentrations of these pollutants are monitored at monitoring stations in each and every city. The purpose of the air quality monitoring station is to track the concentrations of specific air pollutants at different times of the day. Particulate matter is abbreviated as PM.

Particle pollution is another term for it. It is made up of a mix of solid particles in the air like ashes, ash, and soot, as well as liquid droplets. Some particles can be seen with the naked eye, while others require an electron microscope to examine. Particle pollution includes particles as small as PM_{10} and as large as $PM_{2.5}$. Both are inhalable particles with diameters of 10 and 2.5 μm, respectively. Exposure to these particles is harmful to the lungs and heart, as well as causing a variety of respiratory problems.

CO is an odourless, transparent, tasteless flammable gas. This is the most common air pollutant. The largest anthropogenic source of CO is vehicles and fossil fuels. Sulphur dioxide is a poisonous, transparent gas with a strong odour. When sulphuric acid reacts with other chemicals, it produces sulphurous acid and sulphate particles. Humans and industrial waste are the main sources of sulphuric acid. Sulphuric acid is produced when fossil fuels are burned. It infuriates the nose and throat, causing coughing. It causes a tightness in the chest area. When asthmatic individuals are exposed to sulphuric acid, they are more likely to be severely impacted.

Three oxygen atoms make up ozone gas (O_3). Ozone is divided into two categories. One kind forms in the high atmosphere and protects humans from ultraviolet radiation, while another forms at ground level. One of the most hazardous contaminants in the atmosphere is ground-level ozone. Ozone is created by chemical reactions between nitrogen oxides (NO_Xs) and volatile organic molecules. In the presence of sunlight, pollutants released by automobiles, industrial boilers, power plants, refineries, chemical plants, and other sources react. Coughing, throat discomfort, chest pain, and airway inflammation are all side effects of inhaling ozone. When patients with bronchitis, emphysema, or asthma are exposed to it for an extended period of time, they will get a severe infection.

Nitrogen oxide is formed as a result of a reaction between nitrogen and oxygen. They do not react when the temperature is low. But, during high temperature, they react to NO_X which then causes acid rain and smog. The major source of nitrogen oxide is the transport and burning of fossil fuels. Ammonia is a gas with a pungent odour. It com bines with sulphates and nitrates to form $PM_{2.5}$. It is a poisonous gas that when humans inhale for a long time will get affected from cardiovascular disease. It is also found in air, water, and soil. When present in the soil, it leads to solid acidification.

Thus, the hourly air quality dataset with 23 features and 26,282 instances has been developed and was employed to perform exploratory data analysis (Table 3).

3.4 Representation of EDA

Exploratory data analysis is an approach for summarizing the essential characteristics of the data. Exploratory plots such as heat map, boxplot, pair plot, and bar chart can be used to summarize and get overall understanding of the data.

Table 3 Sample air quality dataset

Datetime	2017-07-01T00:00:00	2017-07-01T01:00:00	2018-01-17T11:00:00	2018-01-17T12:00:00	2019-03-07T12:00:00	2019-03-07T13:00:00	2020-01-20T12:00:00
Feels like	29.6	29.6	34.1	34.1	40.6	40.6	36.5
Dew	23	23	23	23	26	26	24
SLP			1012.3		1012.5	1014.3	
Cloud cover	50	50	73.4	50	27.3	27.3	27.3
Visibility	3	3	4.4	5	6	6	8
Conditions	0	0	0	0	0	0	0
Icon	1	1	0	0	0	0	1
$PM_{2.5}$	12	14	85.25	71	69.5	65.5	20.25
PM_{10}	44	38.25	117.5	110	227	223	36
CO	0.56	0.56	1.41	1.27	0.82	0.79	0.89
SO_2	3.45	3.9	6.7	9.25	14.05	24.18	5.32
Ozone	10.12	12.52	48.95	57	67.88	65.83	50.23
NO_X	2.77	2.73	11.17	9.1	5.15	4.18	105.95
NH_3			8.48	7.45	2.88	2.9	13.5
Temp	27.13	27.18	29.75	30.7	33.58	34.25	29
RH	85	85	61.75	57.5	64.25	66	67.75
SR	22	22	502.5	612	735.5	718.25	483.25
BP	756	755.75	761.5	760	756.5	755.25	
AT	33.57	33.52	29.3	29.65	31.15	31.02	31
RF	6.5	1.25	0	0	0	0	0
WS	1.12	1	2.72	2.6	3.55	3.67	3.45
WD	213	208	343.5	344	191.25	179.25	349

(continued)

Table 3 (continued)

Datetime	2017-07-01T00:00:00	2017-07-01T01:00:00	2018-01-17T11:00:00	2018-01-17T12:00:00	2019-03-07T12:00:00	2019-03-07T13:00:00	2020-01-20T12:00:00
AQI	57	57	187	187	120	124	86

Heat Map

Heat map is a visual representation of data present in the dataset. Each and every value is represented as colours. The colours used depend upon the intensity of the data. Strong colours are used to represent high valued data, whereas cool colours are used to denote low values [9]. The principal goal of generating the heat map is to understand the impact of all the values in the dataset in a single form. In this paper, a two-dimensional correlation heat map is generated to identify the influence of features in building an air quality prediction model.

Histogram

Histogram gives a visual overview of continuous data in terms of ranges or bins [10]. The bins are non-overlapped. The general purpose of drawing a histogram is to show the frequency of the occurrence of a feature. Even though histograms look similar to bar charts, they are not the same. Bar charts are used to visualize categorical variables, whereas histograms use continuous numerical variables. The length of the histogram bar denotes the frequency of the occurrence, whereas depth represents the range. In this paper, histograms are generated to identify the distribution of data and the frequency of the occurrence.

Pair Plot Analysis

A pair plot is one of the most effective tools for exploratory data analysis. It is used to identify pairwise relationships in a dataset. The pair plot function creates a grid of axes with a single row on the x axis and a single column on the y axis for each variable in the data. Pair plots are built using a two-figure histogram and a scatterplot. Histograms are used to represent the distribution of a single variable, whereas scatterplots are used to show the relationship between two variables. In this paper, pair plots are generated to visualize the correlation between any two variables in the dataset.

Boxplot Analysis

The boxplot, also known as the box and whiskers plot, is a visual representation of data that uses five numbers to summarize the data: minimum value, maximum value, median, first quartile, and third quartile. Boxplots are used to show the distribution of variables. Boxplots are also used to identify anomalies in the dataset. Outliers can be detected. In this paper, boxplots are generated to see the distribution of meteorological features and pollutant data. Outliers were also identified.

4 Results and Discussion

Exploratory data analysis is performed using Python. The libraries such as matplotlib, seaborn, and sklearn were explored to analyse the meteorological and air pollutant data collected for the period of 3 years from Trivandrum. The air quality dataset

includes 26,544 instances and 23 features. Heat maps are generated to give the visual overview of the entire dataset using different colours. Pair plots were drawn to identify the correlation between meteorological factors and air quality index. Bar charts were drawn to identify the distribution of the data like the minimum value, maximum value, median, first quartile, and third quartile.

4.1 Heat Map

Heat map is generated which gives the visual representation of the entire dataset. They are used to visualize association matrices. It depicts the relationship between the variables in terms of different colours. It is a 24 × 24 matrix since the dataset includes 24 features. The heat map generated is given in Fig. 5. The intensity of the colour indicates the value of the main variable in the corresponding cell range.

From the heat map, it has been identified that dew is positively correlated with feels like, cloud cover, conditions, air temperature, and wind speed. It is negatively correlated with sea level pressure, $PM_{2.5}$, PM_{10}, CO, SO_2, ozone, NO_X, NH_3, and temp. Feels like, visibility, PM_{10}, sulphur dioxide, ozone, sun radiation, air temperature, wind speed, wind direction, and the air quality index are all positively connected with temperature. Sea level pressure, $PM_{2.5}$, PM_{10}, CO, SO_2, ozone, NO_X, NH_3, Temp, and wind direction are positively correlated with AQI whereas feels like, dew, cloud cover, conditions, icon, rainfall, and wind speed are negatively correlated with AQI. There is no correlation between relative humidity, barometric pressure, and AQI. The correlation between all the attributes in the dataset and air quality index is provided in Table 4.

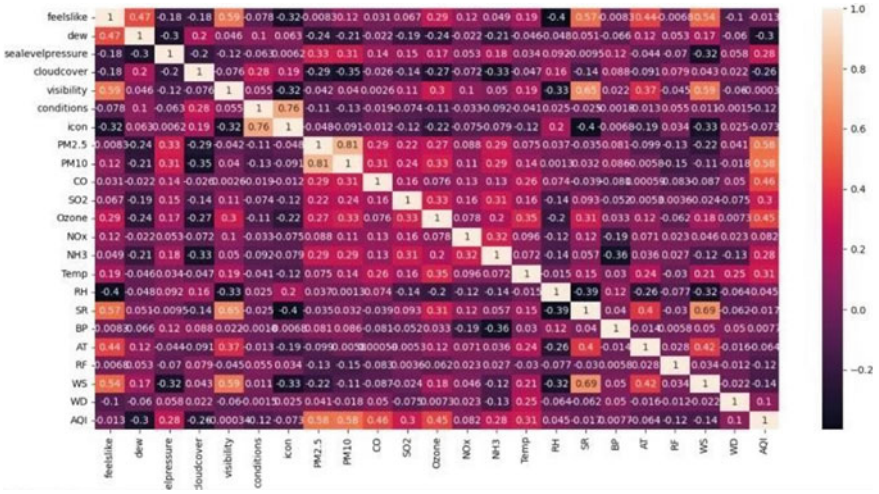

Fig. 5 Heat map depicting the correlation between attributes

Table 4 Correlation between the attributes and air quality index

S. No.	Attribute	Correlation value	Result
1	Feels like	− 0.012688724606251351	Negatively correlated
2	Dew	− 0.3015968201961364	Negatively correlated
3	Sea level pressure	0.28390300278636227	Positively correlated
4	Cloud cover	− 0.2631286566012311	Negatively correlated
5	Visibility	− 0.000343232025569354	Negatively correlated
6	$PM_{2.5}$	0.5756957872056067	Positively correlated
7	PM_{10}	0.5810464548312541	Positively correlated
8	Carbon oxide	0.45764413603029847	Positively correlated
9	Sulphur dioxide	0.30074500670790966	Positively correlated
10	Ozone	0.44706794048480963	Positively correlated
11	Nitrogen oxide	0.18179587794498741	Positively correlated
12	Ammonia	0.28033571115301464	Positively correlated
13	Temperature	0.3119078283575685	Positively correlated
14	Relative humidity	0.045415418707998985	No correlation
15	Solar radiation	− 0.017490877066920537	Negatively correlated
16	Barometric pressure	0.0076884989859714635	No correlation
17	Air temperature	− 0.06382006724829173	Negatively correlated
18	Rainfall	− 0.1207312355884903	Negatively correlated
19	Wind speed	− 0.1365778017395007	Negatively correlated
20	Wind direction	0.10193251948936194	Positively correlated
22	Conditions	− 0.11623631138814557	Negatively correlated
23	Icon	− 0.0733477949698372	Negatively correlated

4.2 Histogram

Histograms are used to identify the distribution and frequency of the data. Histogram generated for the entire dataset is given in Fig. 6. From the analysis, it is identified that amongst the three years, for the maximum number of days, the humidity falls within the range of 65–70. Temperature observed for maximum number of days was between 27 and 30. Dew value ranges from 23 to 24. The minimum value for the humidity is 50, and maximum value is observed as 100. For the maximum number of days, sea level pressure was between 1009 and 1011. Similarly, for maximum number of days, the condition observed was partially cloudy. AQI value ranges from minimum of 20 to maximum value 230.

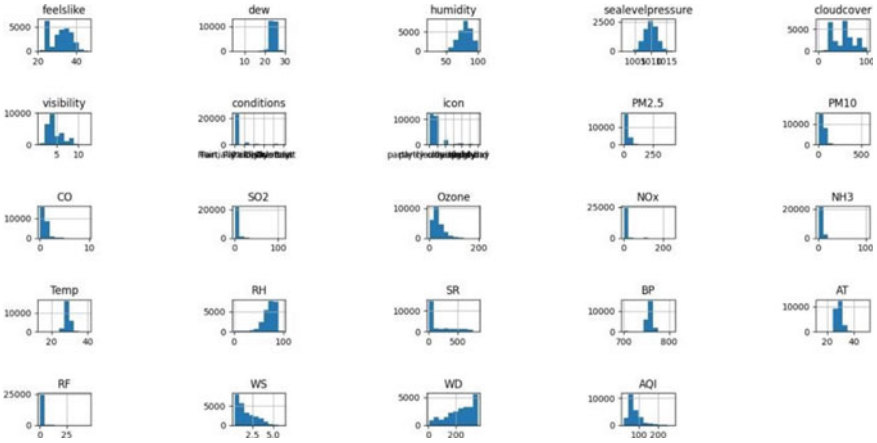

Fig. 6 Histogram depicting frequency distribution between pair of features

4.3 Pair plot

Pair plots are used to identify the correlation between the pair of variables in the dataset. Pair plots are generated taking each and every attribute in x axis and air quality index in the y axis. Relationships between each parameter are visualized using bar graphs and scatterplots [11]. If there are fewer scatter plots, then there is less correlation. The bar graph in the pair plot shows the range of values within which it lies and how many instances lie within the range. The pair plots generated depict that the features sea level pressure, $PM_{2.5}$, PM_{10}, CO, NO_X, NH_3, SO_2, and ozone are positively correlated with air quality index whereas feels like, dew, humidity, wind speed, cloud cover are negatively correlated with air quality index. Pair plots generated were given in Fig. 7

4.4 Boxplot

Boxplot depicts the distribution of data using five number summary. Boxplots are generated for each and every feature to understand their distribution. In addition to that, boxplots also helps to identify the outliers present in the dataset [12]. The attributes such as minimum, maximum, first quartile, median, and third quartile values were determined using boxplots. Amongst the three years the minimum temperature felt was 21 and at the maximum temperature was 36. The minimum value of cloud cover is 0 and the maximum is 100. Minimum temperature observed was 13.5 and maximum temperature was 40.7. The value of sea level pressure ranges between 1001 and 1017. Rainfall ranges from 0 to 45 cm. It has been identified that the features humidity, wind speed, Sea Level Pressure and all pollutant data has

Fig. 7 Pair plot depicting pairwise relationship

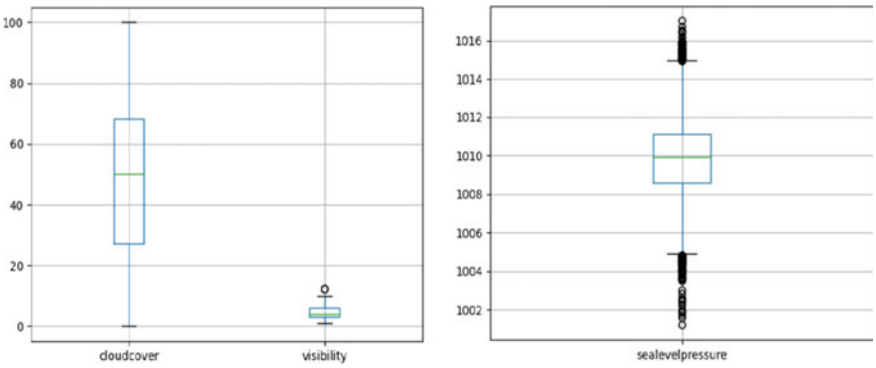

Fig. 8 Boxplot analysis on air quality data

outliers. The boxplot generated is given in the Fig. 8. The five number summary obtained for all the features are provided in Table 5.

4.5 Bar Chart

Bar chart looks similar to histograms but they differ in their purpose. Histograms are drawn to identify the frequency of the occurrence whereas bar charts are used to compare the features in the dataset. From the analysis, it has been identified that maximum temperature 28.925 was felt during the year 2020. The average $PM_{2.5}$ value

Table 5 Five number summary of the air quality data

S. No.	Attribute	Min	25th Percentile	Median	75th Percentile	Max
1	Feels like	21	26	32.4	35.6	46.2
2	Dew	5	23.6	24	25	29
3	Sea level pressure	1001 2	1008.5	1009.9	1011.2	1017
4	Cloud cover	0	27.3	50	68.2	100
5	Visibility	1	3	4	6	12.4
6	$PM_{2.5}$	0.25	12.5	22.75	38.25	418
7	PM_{10}	1	32	48.5	68.75	568.2 5
8	Carbon oxide	0	0.67	0.88	1.15	9.84
9	Sulphur dioxide	0.1	3.32	5.52	8	111.8
10	Ozone	0.03	19.6	30.27	46.03	197.1
11	Nitrogen oxide	0	2.88	5.07	8.38	247.4
12	Ammonia	0.1	2.4	3.98	6.57	102.8
13	Temperature	13.45	28.02	28.77	29.7	40.47
14	Relative humidity	1	65.5	74.5	82.25	100
15	Solar radiation	10.5	22	28	331.75	845.5
16	Barometric pressure	700	753.25	756.25	760.25	807.2 5
17	Air temperature	13.6	27.05	28.75	30.45	49.5
18	Rainfall	0	0	0	0	45.62
19	Wind speed	0.3	0.75	1.33	2.5	6.23
20	Wind direction	5.5	168.5	250.25	316.5	359
21	Conditions	0	2	2	2	6
22	Icon	0	3	4	4	7
23	Air quality index	22	56	68	90	273

is observed as 33.053 during the year 2018. Similarly, the values of other pollutants such as SO_2, PM_{10}, ozone were high during the year 2018.The bar charts generated are given in Fig. 9.

From the EDA made using the visualization aids, it has been identified that sea level pressure, $PM_{2.5}$, PM_{10}, CO, SO_2, ozone, NO_X, NH_3, temp, and wind direction are positively correlated with AQI whereas feels like, dew, cloud cover, conditions, icon, rainfall, and wind speed are negatively correlated with AQI. There is no correlation between Relative Humidity, Barometric Pressure and AQI. The features relative humidity, wind speed, sea level pressure and all pollutant data have outliers. The features that are identified with outliers have to be preprocessed. A wide range of values was observed for all the pollutants such as $PM_{2.5}$, SO_2, ammonia etc., and such attributes have to be standardized using normalization.

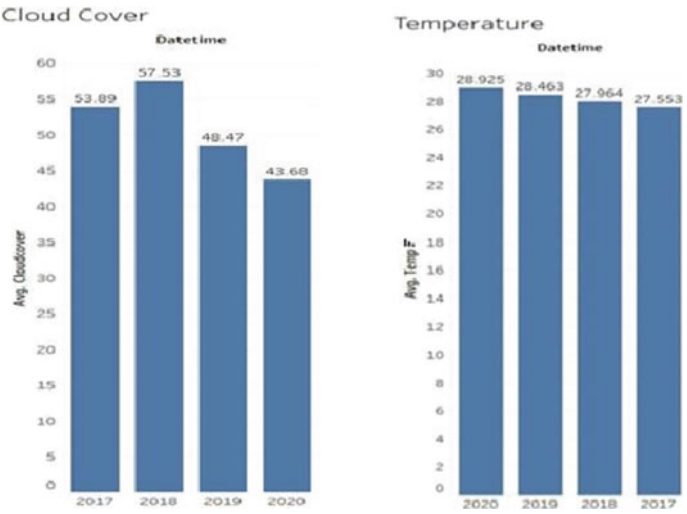

Fig. 9 Bar charts on air quality dataset

The features such as sea level pressure, $PM_{2.5}$, PM_{10}, CO, SO_2, ozone, NO_X, NH_3, temp and wind direction, feels like, dew, cloud cover, conditions, icon, rainfall, and wind speed are either positively or negatively correlated with AQI. These attributes can be used directly for building air quality prediction model. The features relative humidity and barometric pressure has neither positive nor negative correlation, it must not be considered for building the air quality prediction model. The features relative humidity, wind speed, sea level pressure and all pollutant data have outliers. Outlier handling techniques such as trimming, binning, discretization, etc., can be applied on those attributes.

5 Conclusion

In this paper, exploratory data analysis was performed over the air pollutant and meteorological data collected for the period of three years from July 1, 2017 till July 1, 2020. The visualization aids such as heat map, pair plot, histogram, boxplot, and bar charts were generated for the air quality dataset. The attributes that contribute more for predicting the air quality index were identified. The results of the analysis help to get an in-depth understanding of the distribution, association trend in the data. As per the results observed, further preprocessing can be done on the features so as to form a proper dataset. The dataset can further be used to build an air quality prediction model which can provide health alerts for the society.

References

1. Aditya CR, Deshmukh CR, Nayana DK, Vidyavastu PG (2018) Detection and prediction of air pollution using machine learning models
2. Pandey A, Brauer M, Cropper ML, Balakrishnan K, Mathur P, Dey S, Turkgulu B et al. (2020) Health and economic impact of air pollution in the states of India: the global burden of disease study 2019. Lancet Planet Health 2021
3. Mukhiya SK, Ahmed U (2020) Hands-on exploratory data analysis with python. Packt Publishing
4. https://trivandrum.nic.in/en/about-district
5. https://app.cpcbccr.com/ccr/#/caaqm-dashboard-all/caaqm-landing/data
6. https://www.visualcrossing.com/weather-data
7. https://www.kippzonen.com/Knowledge-Center/Theoretical-info/SolarRadiation/Parameters-of-Meteorology
8. Mansouri B, Ebrahimpour M (2011) Monitoring of air quality parameters at different months: a case study from Iran. Cont J Water Air Soil Pollut 2(2):25–31
9. Bartholomeus CM et al. (2015) Feature-expression heat maps—a new visual method to explore complex associations between two variable sets. J Biomed Inf 53:151–161
10. https://www.analyticsvidhya.com/blog/2021/06/exploratory-data-analysis-using-data-visualization-techniques
11. https://www.machinelearningplus.com/plots/python-boxplot/
12. https://medium.com/@jaimejcheng/data-exploration-and-visualization-with-sea-born-pair-plots-40e6d3450f6d